Solar Energy Materials & Solar Cells 263 (2023) 112559

Contents lists available at ScienceDirect

Solar Energy Materials and Solar Cells

journal homepage: www.elsevier.com/locate/solmat

Machine learning assisted representative period selection as input to modelling of field degradation in photovoltaic modules

Gaute Otnes [*], Dag Lindholm, Hallvard Fjær, Pernille Seljom, Sean Erik Foss [1]

Institute for Energy Technology, 2007, Kjeller, Norway

ARTICLE INFO

Keywords:
PV modules
Numerical modelling
Field degradation
Machine learning
K-means clustering

ABSTRACT

Sustainable and affordable solar energy production is critically dependent on the ability of photovoltaic (PV) modules to perform reliably in the outdoor environment over several decades. A suite of accelerated stress tests is central in the efforts to qualify the reliability of new module architectures and materials, which are evolving at a rapid pace. Complimentary to these, mathematical modelling of PV module degradation phenomena constitutes a useful set of tools that has become increasingly relevant in recent years. For certain degradation phenomena, modelling is best solved by numerical methods such as using the finite element modelling framework. These numerical methods can be computationally expensive, thus the input of longer timeseries of outdoor weather data can come at a high computational cost and require significant model simplifications. To alleviate this problem, we here present the use of an unsupervised machine learning algorithm to select representative periods of historical outdoor weather data (or derivatives thereof). We exemplify the approach by selecting representative daily cell temperature profiles, instead of using the full timeseries, as input to modelling of thermomechanical fatigue in soldered ribbon cell interconnects. We explore how a subset of representative daily temperature profiles can reproduce key characteristics of the overall distribution of a multi-year dataset, as well as differences between datasets for sites in different climates. Such representative datasets could be used as input to complex PV module models and drastically reduce the computational costs.

1. Introduction

The reliability of the photovoltaic (PV) modules is central to the profitability of any PV plant. Stressors present in the outdoor environment, such as temperature, temperature variations, ultraviolet light, humidity, and various mechanical loads, can over time provoke failures and/or performance degradation [1]. While there is evidence that PV modules can maintain performance to and beyond the typical limits in current performance warranties [2–4] the observed performance loss rates are strongly bill-of-material (BOM) and module design dependent [5]. At the same time, module designs and BOMs are constantly experiencing incremental changes to cut cost and improve efficiency. Hence, the currently available products will always deviate somewhat from the products with a proven field-record, and even more substantially so from products where long field experience exist [6].

To qualify new module designs before field data is available, a suite of accelerated stress tests is therefore typically employed. These tests have been developed over decades to provoke known types of infant field degradation at short testing times [7] thus providing an efficient means to quickly uncover quality issues in the design. However, current accelerated stress tests are not lifetime tests, thus a quantitative correlation between accelerated test results and outdoor performance loss or failure does not exist. Also, the tests have typically not been designed to provoke wear-out or end-of-life degradation modes. And further, there is no guarantee that the current tests can pick up new degradation modes arising when introducing new materials and module designs [8].

Many strategies are explored to improve the conventional accelerated stress testing regime. These include new test strategies such as testing with stressors combined and in sequence [9–11] with adapted loads for specific climates [12,13] or on material level [14–16]. Complimentary to this, interesting possibilities lie in the mathematical modelling of the degradation processes. Such models enable more efficient module development, where reliability concerns can be explored and accounted for already at the design stage. If well developed and validated, such models should ultimately also give the ability to do service life prediction modelling [17] providing acceleration factors

[*] Corresponding author.
E-mail address: gaute.otnes@ife.no (G. Otnes).

[1] Present affiliation: Over Easy Solar AS, 0252 Oslo, Norway.

https://doi.org/10.1016/j.solmat.2023.112559
Received 8 June 2023; Received in revised form 12 September 2023; Accepted 14 September 2023
Available online 25 September 2023

Table 1
Specifics of the six weather datasets used to model cell temperatures.

Location	Data-source reference	Time-period (date format- DD.MM.YY)	Original time-resolution [min]			Sensor height [m]			Latitude [°N] Longitude [°] Altitude [m]	Köppen-Geiger climate zone (Found by use of [54])	% of timestamps missing at least one parameter	% of days replaced (days with gaps >2 h)
			T_{amb}	W_s	GHI	T_{amb}	W_s	GHI				
Oslo, Norway	[55]	01.03.16–28.02.21	60	10	1	2	10	NA	59.94 10.72 77	Dfc-Subarctic Climate	6.7	2.2
Freiburg im Breisgau, Germany	[56]	01.01.10–31.12.19	10	10	10	2	12	6	48.02 7.83 236	Cfb-Temperate Oceanic Climate	0.36	0.68
Tucson, AZ, USA	[57]	01.01.13–31.12.20	5	5	5	1.5	1.5	1.5	32.24 −111.17 844	Bsh-Hot Semi-arid Climate	0.04	0.21
Everglades, FL, USA	[57]	01.01.13–31.12.20	5	5	5	1.5	1.5	1.5	25.90 −81.3 1.5	Am-Tropical Monsoon	0.16	0.27
Fallbrooks, CA, USA	[57]	01.01.14–31.12.20	5	5	5	1.5	1.5	1.5	33.44 −117.19 344	Csa-Mediterranean hot summer	0.07	0.19
Sioux Falls, SD, USA	[57]	01.11.12–31.10.19	5	5	5	1.5	1.5	1.5	43.73 −96.62 497	Dfa-Hot Summer Continental Climate	0.51	0.90

between accelerated stress tests and outdoor exposure under various conditions.

Building mathematical models of the degradation processes is however a complicated task [18]. To capture a full degradation pathway typically involves several modelling steps: The starting input is the macroclimatic conditions, the module architecture, and the system configuration, which is translated into microclimatic/local loads and stress levels. These are subsequently used as input to chemical and/or physical models of the degradation in one or more of the module materials. Finally, this degradation needs to be translated into a performance loss. The complexity and possible simplifications in each of these steps will vary significantly depending on the degradation mode in question. Considering multi-dimensional and interlinked reaction paths will add another level of complexity. It should also be noted that any model aiming for service life prediction capabilities must additionally take into consideration both variability in BOMs and the manufacturing process.

As mentioned in the previous paragraph, an important input to model field degradation is information about macroclimatic conditions, thus some form of weather data. For some modelling purposes and/or degradation processes, simplifications from temporally varying parameters can be made, e.g. by using weighted average temperatures [19] or humidities [20] as input to analytical degradation models. For other processes or purposes, it is instead desirable to input weather data time-series into more complex numerical models, which can come at a high computational cost. To illustrate this problem, we will in the following look specifically at the modelling of thermomechanical fatigue in soldered ribbon cell interconnects. However, we note that the problem holds relevance also for modelling of other degradation modes, such as cell crack effects during cyclic mechanical loads [21–23], crack development in backsheets driven by thermomechanical stress [8] or coupled hygro-thermo-mechanical effects [24].

Modelling of PV modules by use of the finite element method (FEM) has been an active research field over the last 10–15 years. An instructive overview of the topic can be found in a recent review by Nivelle et al. [25]. PV modules represent a challenging modelling problem due to their high aspect ratio. Depending on the modelling goal, the model might need to capture phenomena on length scales ranging from 10^{-5} m (busbar and solder thickness) to 10° m (module size). To achieve this, multi-scale modelling approaches can be applied [26–32]. Type of computational domain (e.g. 2D or 3D), choice of boundary conditions (e.g. symmetry), constitutive material models, multi-scale approaches etc. influence on the complexity of the problem and hence computational demands [25]. Simplifications does however come with a risk of lower accuracy and less relevant results. Further coupling mechanical phenomena to chemical changes in a multi-physics modelling approach [28] will give additional complexity.

For thermomechanical stress on the interconnects, several FEM modelling studies have investigated its dependence on module BOM, manufacturing and design [26,31,33–41]. For such studies, the model is typically run through the temperature profile of the IEC 61215 thermal cycling test [42] and the computed damage is compared between cases. Studies have also been performed to explore the thermal cycling temperature profile itself, and how it can be modified to e.g. speed up testing [43]. Computing damage for actual outdoor temperature profiles is computationally demanding, however, making a comparison between

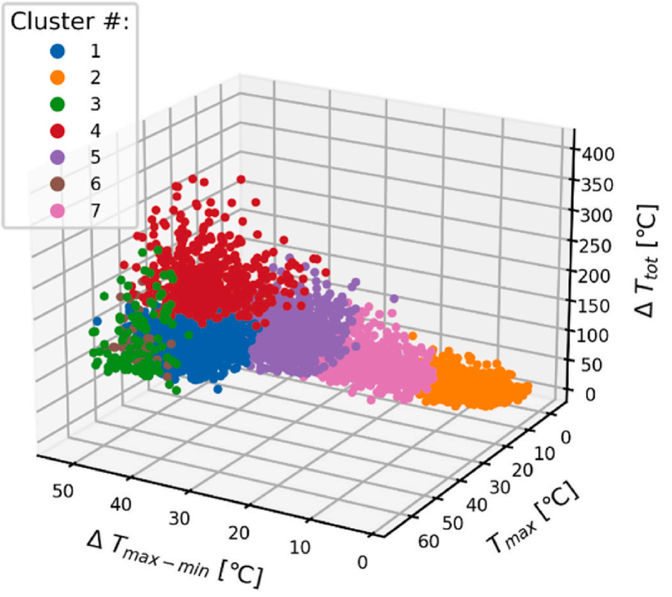

Fig. 1. Each of the days in the ten year (3652 days) dataset for Freiburg im Breisgau plotted according to the three parameters T_{max}, $\Delta T_{max\text{-}min}$, and ΔT_{tot} (see main text, section 3, for parameter definition). The color-coding of the datapoints place each day into one out of seven clusters, as determined by the k-means clustering algorithm. Note that days from cluster number (#)3 and #6 are largely interspersed in this plot, as the main parameter separating these two clusters is N_{rev}.

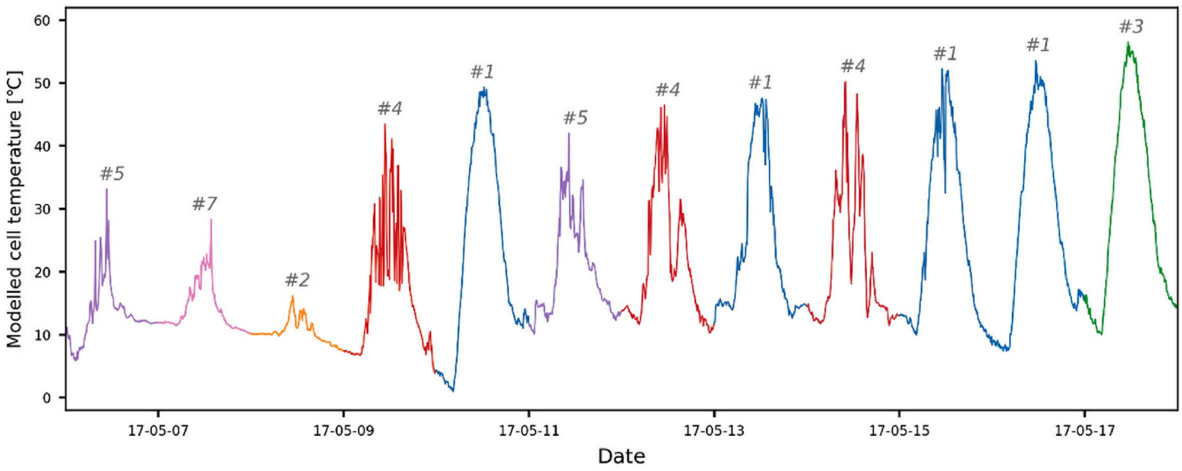

Fig. 2. An example period from May 2017, showing twelve daily temperature profiles color coded equivalently to Fig. 1, based on their assignment to different clusters by the k-means clustering algorithm. For clarity, the cluster number (#) to which each daily temperature profile is assigned is displayed in grey color above the respective curve.

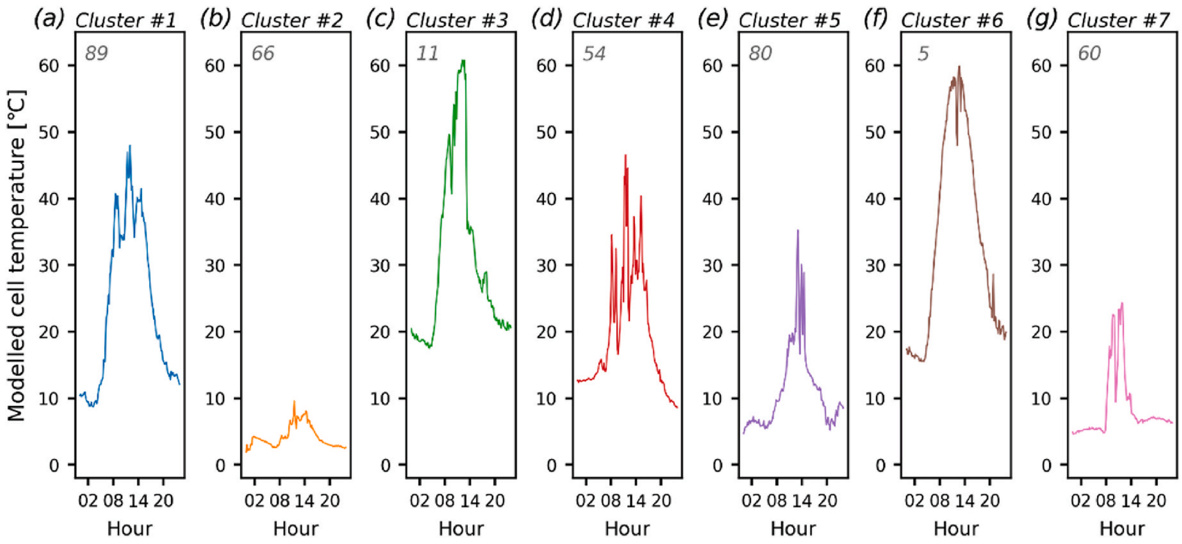

Fig. 3. The representative daily temperature profile selected for each of the seven clusters shown in Fig. 1. The representative day was selected as the medoid datapoint for each cluster. In the top left corner of each sub-figure, the average count of each cluster per year is specified in grey color.

computed damage during indoor thermal cycling and outdoor field exposure difficult. To overcome this issue, a few different approaches have been taken. First, simpler 2D models were used to compute accumulated damage during full years of outdoor temperature variations for various locations [43,44]. The 2D models focused on stresses in the solder layer and enabled low computational demands but might be less capable of studying stresses in the ribbon and of capturing long-range effects such as the influence of mounting conditions. Second, statistical analysis of outdoor temperature profiles has been used to generate an "average" thermal cycle for a given location [45,46]. Such a cycle is easy to model but does not contain e.g. temperature fluctuations happening on short timescales and might not be easy to determine in locations of large seasonal variations. Third, numerous studies have selected and modelled a few days from a given location, representing certain conditions such as "sunny and hot", "cloudy and hot", "summer day" or "winter day" [33,40,44,47,48]. The approach of only modelling certain days intuitively makes a lot of sense. Clearly there are many days in a year that have quite similar temperature profiles and hence result in similar thermomechanical damage; it should not be necessary to model them all. However, the days to model have so far been manually selected in a heuristic manner. This makes it difficult to judge and ensure the representability of the selected days for the site. Further, it does not allow objective and consistent comparisons between sites, climates and/or system configurations. In sum, there is a clear need for an approach that can allow for modelling of realistic outdoor temperature profiles at acceptable time-resolution, while using complex models, and in a way that is reproducible and comparable between climates, sites and/or system configurations.

In this work, we use an unsupervised machine learning algorithm to select a set of representative daily temperature profiles from historical timeseries for various locations. Such algorithms have been used to select representative historical periods of input data in other modelling disciplines, such as for long-term energy system modelling [49–51] and in finance [52]. Various algorithms exist to select periods which together give a good statistical representation of the overall variety of the dataset. A common approach is to use clustering algorithms [53], i.e. machine learning techniques that group datapoints. In this work, we use k-means clustering, one of the simplest and most common clustering techniques. We use this algorithm to select representative daily PV cell temperature profiles for six different locations in different climate zones. We show how the algorithm can group similar temperature profiles and select a representative temperature profile for each group. Further, a

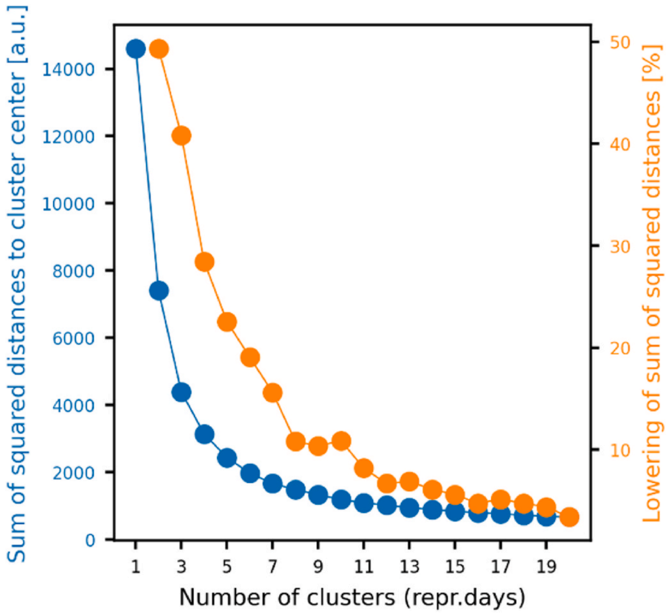

Fig. 4. An elbow plot for the Freiburg im Breisgau ten-year dataset. The sum of squared distances from all datapoints to their cluster center is plotted as a function of the number of clusters (blue, left y-axis), as well as the percentage-wise lowering of this parameter when adding an additional cluster (orange, right y-axis).

yearly frequency for each representative profile is obtained to enable the creation of a full-year representative dataset, based on only a subset of days. We study how such a representative dataset can reproduce key characteristics of the original dataset, as well as differences between sites located in different climates. Such representative datasets, consisting of a low number of temperature profiles, could be used as input to complex PV module models and drastically reduce the computational efforts.

2. Input data and methods

2.1. Processing and quality control of the weather data sets

As input to FEM modelling of thermomechanical fatigue in interconnects, cell temperature data is the relevant microclimatic stressor. For the current work, we will model such data from weather data sets. To evaluate our approach, we have chosen weather station datasets gathered for selected locations in a variety of different climate zones from open databases. Information on the datasets is summarized in Table 1.

The time-period for each dataset was chosen to get a time-series of minimum 5 years, containing only full years, and with good data-availability for all relevant weather parameters. The relevant weather parameters for modelling of cell temperature are ambient temperature (T_{amb}), wind-speed (W_s) and global horizontal irradiance (*GHI*). All datasets contained quality information for each timestamp, based on which all data flagged as erroneous was removed. Thereafter, the percentage of datapoints missing at least one parameter was determined (see Table 1). Missing data was replaced as follows: For gaps longer than 2 hours, data for the entire day or set of days were replaced by data from the same day(s) in the following year. The percentage of days that were replaced in this way is shown in the last column of Table 1. All gaps shorter than 2 hours were filled by linear interpolation.

Based on the time-resolution of the available datasets and the need to keep the computational efforts reasonable while at the same time capturing relevant temperature changes, a 10 minute time-resolution was chosen as a common standard for all datasets. Higher time-resolution datasets were down-sampled to a 10 minute time-resolution by removing excess datapoints. For the Oslo location, the 1 hour T_{amb}-data were up-sampled to 10 minute time-resolution; the variation in temperature between the original measurement points were set based on

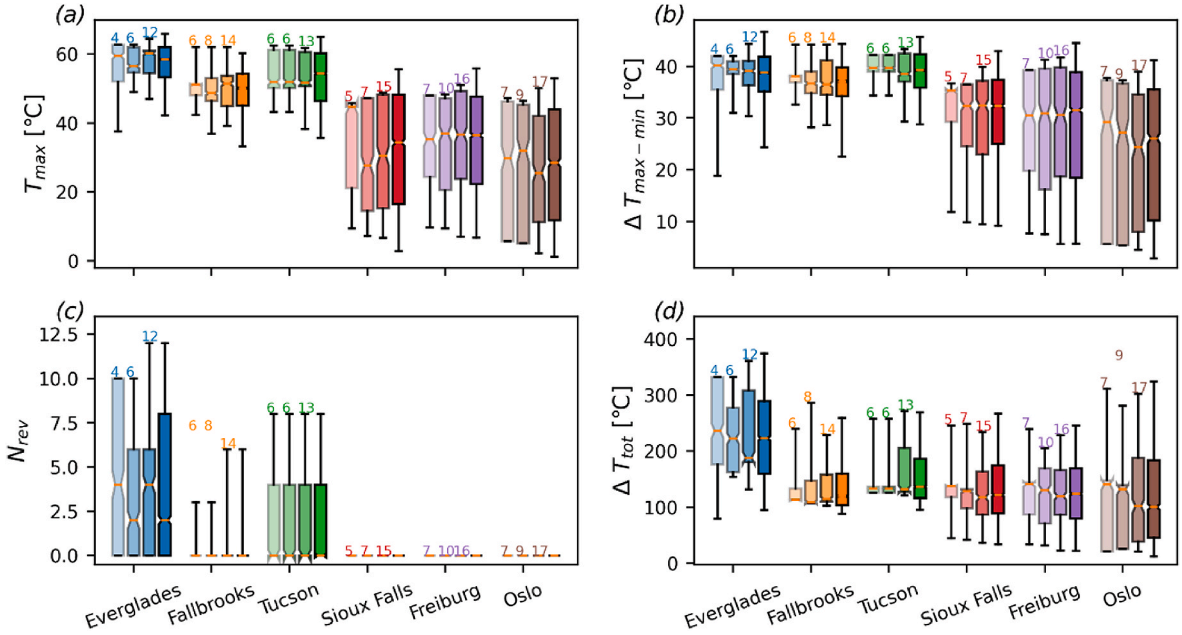

Fig. 5. For each of the six locations, boxplots are plotted for (a) T_{max}, (b) $\Delta T_{max\text{-}min}$, (c) N_{rev}, and (d) ΔT_{tot} (see main text, section 3, for parameter definition). For each location, boxplots are shown for representative datasets constructed by using a 15, 10 and 5% threshold in the elbow plot to select the number of representative days, as well as for the original dataset (boxes from left to right for each location respectively, with increasing color saturation). The number of representative days making up each dataset is shown by the numbers above the upper whisker. The features of the boxplot can be interpreted as follows: orange line shows the median, the notch represents the confidence interval around the median, the box extends from the first to the third quartile of the data, while the whiskers give the 5th and 95th percentile.

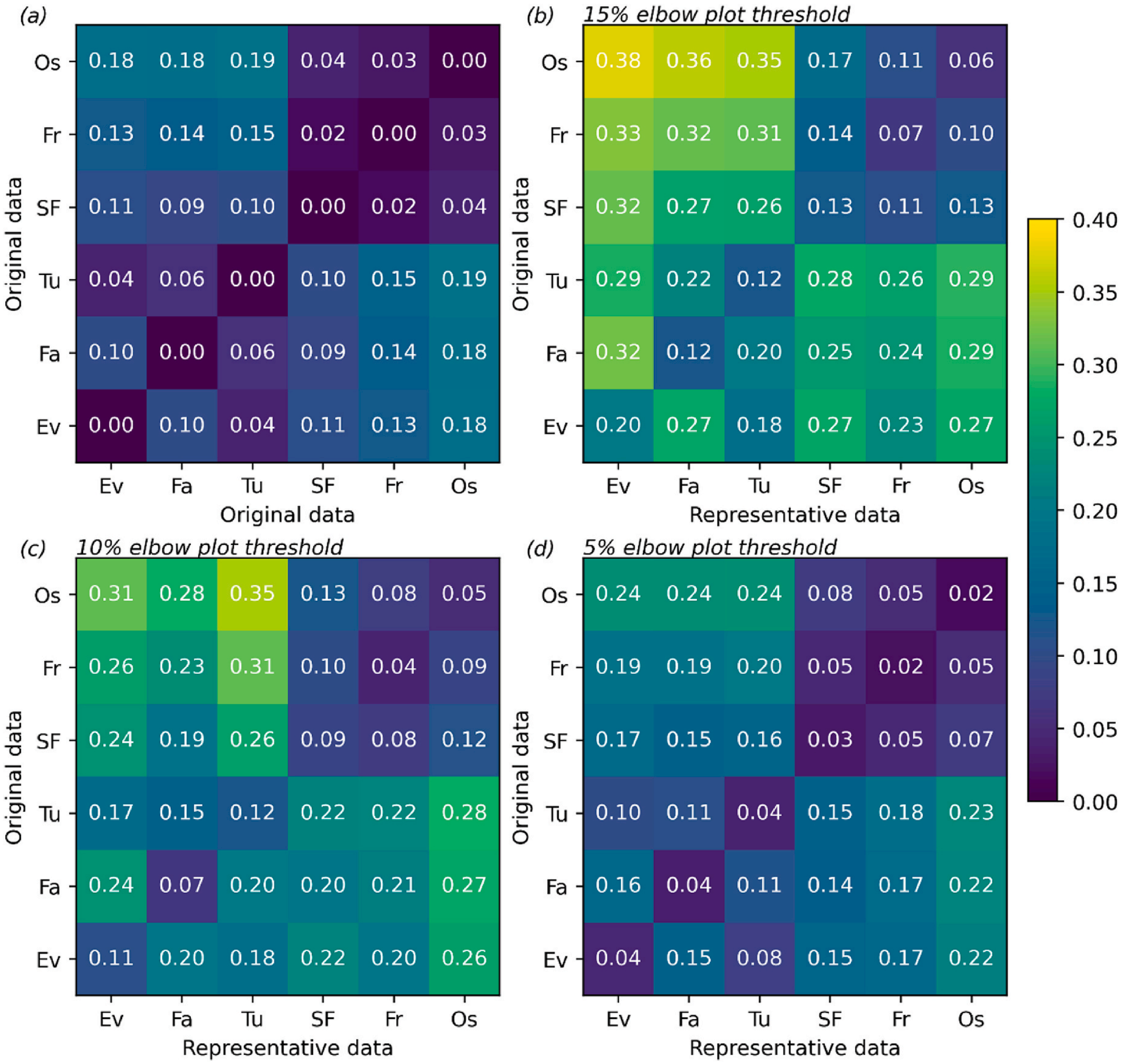

Fig. 6. Heatmaps where energy distances [a.u.] are calculated between distributions for the six different locations (Os- Oslo, Fr- Freiburg, SF- Sioux Falls, Tu- Tucson, Fa- Fallbrooks, Ev- Everglades). The original datasets (y-axis) are compared to (a) the original dataset, as well as representative datasets built by applying an elbow plot threshold of (b) 15%, (c) 10% and (d) 5% (x-axis).

10 minute time-resolution temperature data available from a measurement station located ~5 km away. For the locations where the wind sensor height differed from the standard meteorological height of 10 meter, Hellmann's equation (as implemented in equation 2 in Ref. [58]) was applied to estimate wind speed at 10 meter height from the measured values.

2.2. Modelling of cell temperatures from weather data sets

From the sets of weather data described above, cell temperatures were modelled by help of the python package pvlib, version 0.8.0 [59]. First, the GHI values were transposed to plane-of-array (POA) irradiance at an optimal fixed tilt for the latitude, selected by Eq. 12 of [60]. Thereafter, the POA irradiance (E_{POA}, W/m^2), T_{amb} (°C), and W_s (m/s) was used to model the cell temperature (T_c, °C), by use of the Sandia (King) model [61], as given in Equation (1).

$$T_c = E_{POA} \times \exp(a + b \times W_s) + T_{amb} + \frac{E_{POA}}{E_0} \times \Delta T, \qquad \text{Equation 1}$$

In Equation (1), a, b, and ΔT are empirically determined coefficients for a "glass/cell/polymer sheet" module type in an open rack mount; −3.56, −0.075 and 3 °C, respectively [61]. E_0 is a reference irradiance of 1000 W/m^2.

2.3. Selection of representative periods

A first question when selecting representative periods is their desired length; one could imagine selecting representative hours, representative days or even representative weeks [50,62]. For the thermomechanical fatigue of interconnects, temperature variations ranging from diurnal to sub-hourly timescales are important, while variations on longer timescales, e.g. weeks and months, are less important. Therefore, we aim at selecting representative daily temperature profiles with a high time-resolution, capturing both diurnal and sub-hourly temperature variations.

To select the representative days we use k-means clustering. In short, the k-means algorithm first initializes a given number of center-points and assigns all datapoints to clusters based on their nearest center-point. From the resulting clusters, new center-points are calculated, and then all datapoints are reassigned to new clusters based on the new center-points. This process continues in an iterative fashion until a convergence criterion is met. In this work, the k-means clustering was

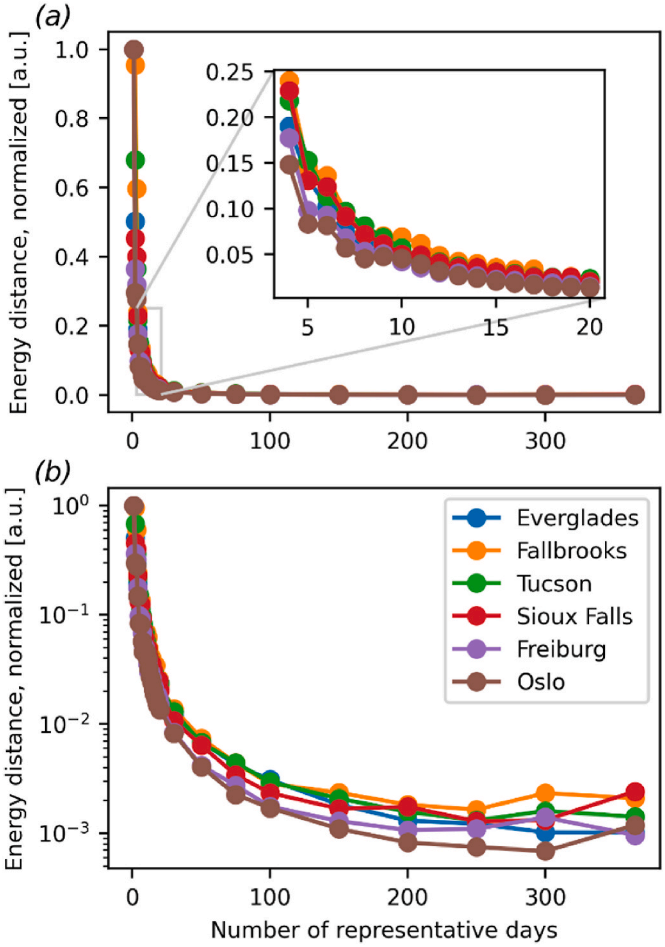

Fig. 7. Plots of the normalized energy distance between the distribution of temperature parameters in the original dataset and in the representative dataset, as a function of number of representative days used to build the representative dataset. Plotted using a linear (a) and logarithmic (b) y-axis.

performed through an implementation for python from scikit-learn (*KMeans*, with default settings) [63]. Before clustering, the data was scaled by help of the StandardScaler from scikit-learn (*fit_transform*, with default settings) [64] to give all parameters equal weight.

2.4. Comparing parameter distributions

To compare the parameter distributions for different datasets (discussed further in the Results and Discussion), the energy distance metric was used [65] as implemented for python in the dcor project, v0.5.6 (*energy_distance*, with default settings) [66]. The energy distance was calculated for scaled distributions, as described for the k-means clustering procedure (section 2.3).

3. Results and Discussion

K-means clustering is a method to group datapoints. In our example case, we want to use it to group daily temperature profiles. Thus, we need to represent each daily temperature profile by a set of parameters to be used as input to the k-means clustering algorithm. In selecting the set of parameters, we base ourselves largely on a study by Bosco et al. [43]. They found that an empirical analytical expression could be used to calculate solder fatigue damage in agreement with FEM-simulations if provided the following input for a given location: the mean daily maximum cell temperature, the mean daily maximum cell temperature change, and a temperature reversal term describing the number of times the temperature history increases or decreases across a reversal temperature, T_{rev}. By fitting to FEM-simulated data, T_{rev} was found to be 56.4 °C, thus this parameter is related to temperature loading cycles happening at fairly high temperatures. Based on these results, we describe each daily temperature profile by the maximum daily cell temperature, T_{max}, the difference between daily maximum and minimum temperature, $\Delta T_{max\text{-}min}$, and the number of temperature reversals around T_{rev}, N_{rev}. In addition, we add a fourth parameter: the total summed temperature change, ΔT_{tot}. This fourth parameter is added to distinguish days of different degree of changing irradiance conditions, also when the ambient temperature and/or irradiance is not high enough to reach T_{rev}.

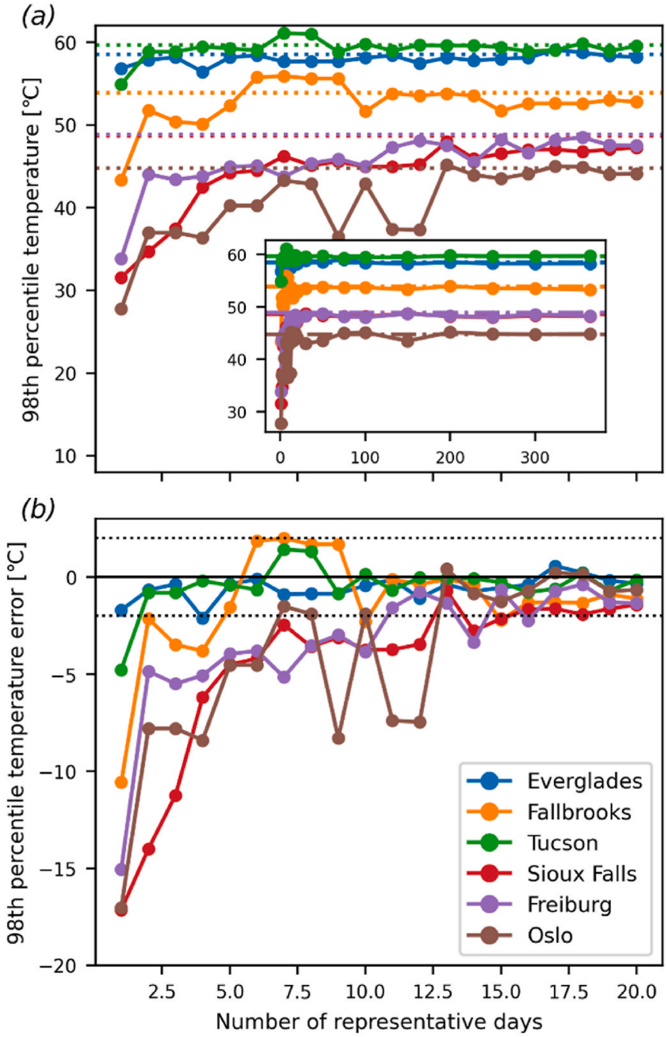

Fig. 8. (a) The 98th percentile cell temperature of the representative datasets plotted as a function of number of representative days. The dotted lines show the 98th percentile temperature for the original datasets. The inset show the same data, for a larger range of representative days used. (b) The difference between the 98th percentile cell temperature of the original datasets to the representative dataset plotted as a function of number of representative days. The black dotted lines indicate a +2 and −2 °C difference.

We note that improvements might be possible in which parameters are used to describe the daily temperature profiles. For example, using a rainflow counting algorithm [44] might be beneficial compared to the use of ΔT_{tot}. Also, Bosco et al. established their empirical model based on a specific module design and solder material [43] and adjustments e.g. to T_{rev} might be needed for other module architectures. However, an in-depth study of the parameter selection is beyond the scope of this paper. Further, we note that work needs to be done to establish appropriate selection parameters for other types of input data and/or degradation modes, such as e.g. wind load patterns or thermomechanical fatigue of polymer materials. Such work might need to be done in an iterative fashion, where modelling of an initial set of representative periods selected based on best guess criteria would give guidance to subsequent refinement of the selection parameters.

After describing all daily temperature profiles by the set of four parameters, each dataset is run through the k-means clustering algorithm. To illustrate the outcome, all days in a ten year dataset (3652 days) for the location of Freiburg im Breisgau is plotted in Fig. 1 as a function of the three parameters T_{max}, $\Delta T_{max\text{-}min}$, and ΔT_{tot}. Each datapoint is color-coded according to the cluster assignment done by the k-means algorithm, into one of seven different clusters. In Fig. 2, the cell temperature profile over a period of twelve days in May 2017 is plotted as an illustration, where the color coding of each daily temperature profile corresponds to that in Fig. 1, based on their assignment to different clusters by the k-means clustering algorithm. The cluster number to which each daily profile is assigned is also displayed above the respective curve. Qualitatively, we observe that similar temperature profiles have been placed in the same cluster. For each cluster, a representative daily temperature profile is chosen as the medoid datapoint, i.e. the datapoint closest to the center of the cluster. These representative profiles are shown in Fig. 3, with the link between the profiles in Figs. 2 and 3 again given by the color of the curves. In the top left corner of Fig. 3 subfigures, the average frequency per year for each profile is written in grey, based on the number of datapoints in each cluster. Thus, the set of representative days and associated frequencies can be used as input to modelling representing a typical year for the location. How well such a dataset can represent the original dataset will be discussed shortly, but first we need to look at how we should choose the number of representative days.

The k-means clustering algorithm (and many other clustering algorithms) does not inherently decide the number of clusters it outputs. The cluster number is instead an input parameter often decided through the inspection of elbow plots. Fig. 4 shows an example elbow plot for the Freiburg im Breisgau dataset, where the sum of squared distances from all datapoints to their closest cluster center is plotted as a function of number of clusters (blue). In a dataset containing an underlying structure or grouping of the datapoints, a clear “elbow” can often be identified in such plots. At the elbow, it is clear that adding an additional cluster will not lead to an improvement that is worth the added cost, and it might also lead to overfitting. In our data however, such an underlying structure or grouping does not exist. Hence, it will be a somewhat subjective decision to decide the point where the diminishing added return is no longer worth the additional cost, which will also be dependent on the purpose, model complexity and the needed level of accuracy. However, to allow for comparative analysis between sites and/or system configurations, we seek a way of selecting the number of clusters which is comparable between datasets. To this end, we look at the corresponding lowering of the sum of squared distances when increasing the cluster number, also plotted in Fig. 4 (orange). When comparing different datasets, we propose to choose a cluster number for all datasets where further increase in the cluster number will not lead to a percentage-wise lowering above a certain threshold. For the example shown in Figs. 1–3, the Freiburg im Breisgau dataset was treated with a 15% threshold, which lead to a selection of seven representative days. A lower threshold should be chosen if larger accuracy is needed, but the threshold should be kept the same when comparing datasets.

At this point, we are able to outline the procedure to select representative periods and how it might be used, using our example of thermomechanical interconnect modelling: First, a suitable representative period, and the parameters to describe the relevant weather data must be selected. In the example case, daily temperature profiles were selected as the most relevant time period, and then described by four different parameters (T_{max}, $\Delta T_{max\text{-}min}$, N_{rev}, and ΔT_{tot}). Second, a suitable number of representative periods need to be set, depending on the required accuracy or other needs. In the example case, we set the number to seven for Freiburg im Breisgau, based on an elbow plot as described above. Third, the clustering algorithm is run. In the example case, the clustering algorithm groups all periods into separate clusters (Fig. 1) and extracts a representative profile as well as average frequency per year for each of the cluster types (Fig. 3). Fourth, this is used as input to the modelling. In the example case, the damage associated with the computed thermomechanical strain (or plastic work) for each of the representative profiles (Fig. 3) would be modelled by the FEM model. Then, to estimate the total interconnect damage generated during one year in the location, the resulting damage for each of the representative profiles would be multiplied by the average frequency per year for each cluster, (e.g. 55 for cluster 0 (Figs. 3a), 49 for cluster 1 (Fig. 3b), etc) and then summed. This estimation of damage could be compared to the damage estimates computed in an equivalent manner for other locations or mounting configurations, as well as for an accelerated thermal cycling profile. For example, climate specific modelling could be done as in Ref. [43] but while employing more complex 3D multi-scale models.

With the method to select the representative days established, we now seek to understand how well these days represent the full dataset. In Fig. 5, the full-year dataset built by the representative days is compared to the original dataset for each of the six different locations/climates. The comparison is done by looking at boxplots for each of the four parameters T_{max}, $\Delta T_{max\text{-}min}$, N_{rev}, and ΔT_{tot}. For each location, the three leftmost boxplots of increasing color saturation (moving left to right) show the distribution for representative datasets built from a varying number of representative days, where the number of days is set by a 15, 10 and 5% threshold in the elbow plot, as discussed above. The resulting number of days in each case is placed above the corresponding upper whisker. The rightmost boxplot in saturated color is showing the distribution of the original dataset, consisting of between 1826 and 3652 days depending on location. Qualitatively, we observe that even when using a low number of representative days, in the range of 5–20, key statistical features of the original dataset, such as the median and the 25th and 75th percentile, is well reproduced. Reproduction of the 5th and 95th percentiles is however more challenging for some of the parameters. To make a representative dataset which better captures such aspects of the original dataset, other methods to cluster the dataset and/or select representative days might have advantages and should be further explored.

To be useful, the representative dataset should not only reproduce key features of the original dataset, but also reproduce significant differences and similarities between datasets. In our case, it should reproduce differences and similarities in temperature variation between e.g. different sites or mounting configurations. Again, the plots in Fig. 5 show how key differences between the different datasets are reproduced, such as differences in median or interquartile range. For a more quantitative comparison, Fig. 6 show heatmaps of the distribution of energy distances calculated between different datasets. The energy distance is a metric that characterizes similarity of distributions [65]. It was calculated between pairs of distributions of the four parameters, as described in the section 2.4. In Fig. 6a, the original datasets are compared to each other. As expected, we observe smaller energy distances (more similar distributions) within the group of warmer (Tucson, Everglades, Fallbrooks; lower left corner) and the group of colder (Sioux Falls, Freiburg im Breisgau, Oslo; upper right corner) locations, than when comparing warm and cold locations (upper left or lower right corner). It is also clear that e.g. Sioux Falls is more similar to the group of warmer locations than Oslo. Fig. 6b compares instead the original datasets to representative datasets selected by use of the 15% elbow plot threshold, thus datasets built by 4–7 days. Already, we observe that key features of the original heatmap, such as those discussed above, is reproduced. As we decrease the elbow plot threshold from 15% (Fig. 6b) to 10% (Fig. 6c) to 5% (Fig. 6d), the overall energy distance between distributions decrease, and the similarity to the original heatmap (Fig. 6a) becomes clearer.

There will always be a trade-off between accuracy and computational cost. To study what might be gained in going to a higher number of representative days, Fig. 7 show a plot of the energy distance between the representative and original dataset as a function of number of representative days, up to a full year. We observe that when increasing the number of representative days from one and onwards, the energy distance falls rapidly in the beginning. As the number of representative days increases however, the energy distance decrease is quickly slowing down. Eventually, there is no significant gain in increasing the number of representative days. It is clear that a full year of data, such as the use of a typical meteorological year [13] is not needed to capture the key features of the original datasets for this example case.

As another metric to evaluate the representative datasets, we look at the 98th percentile cell temperature, $T_{c,98}$. The 98th percentile module temperature (which is closely linked to $T_{c,98}$) has been adopted in IEC TS 63126 [13,67] as a metric to evaluate whether deployment in a certain location and system configuration warrants modified qualification and safety testing of the module. In Fig. 8a, $T_{c,98}$ of the representative datasets is plotted as a function of number of representative days used to build the dataset, and compared to $T_{c,98}$ of the original dataset. In Fig. 8b, instead the difference in $T_{c,98}$ between the original and representative datasets is plotted. At a low number of representative days, the $T_{c,98}$ is too low, but quickly increases towards the original dataset value. For all locations, the value stabilizes within 2 °C of the true value at a number of representative days lower than 15. In line with Figs. 5 and 6, the differences between the datasets are well reproduced.

4. Summary and conclusions

We have shown how an unsupervised machine learning technique, the k-means clustering algorithm, can be used to select representative historical periods of weather data (or derivatives thereof) as input to computationally intensive modelling of field degradation in PV modules. We have exemplified the approach by selecting representative daily cell temperature profiles for six different locations in different climate zones, to be used as input to FEM modelling of thermo-mechanical fatigue in cell interconnects. The method is capable of grouping temperature profiles based on a set of descriptive parameters and selects a representative profile for each group. From this, a representative yearly dataset can be built. We have studied how such a dataset compares to the original dataset and shown that key statistical features are reproduced. It seems clear that even when using only a low number of representative daily profiles, in the range of 5–20, key characteristics of the temperature variations at a site, as well as differences between sites, are reproduced. Using such representative days instead of a full year of data, would cut the computational cost by a factor of 20–70. Lastly, we have presented an approach to allow representative period selection for different sites or system configurations in a comparable way, through thresholding in elbow plots.

We believe that the presented approach constitutes a useful addition to the toolbox for the modelling of PV modules, especially as the modelling becomes more complex, incorporating both multi-scale and multi-physics effects. Further work will be needed to e.g. couple the presented work to numerical modelling and establish the full modelling pipeline proposed, assess suitability and develop appropriate selection parameters for other degradation modes, and assess the possibility of improved clustering algorithms.

CRediT authorship contribution statement

Gaute Otnes: Writing – review & editing, Writing – original draft, Visualization, Software, Methodology, Investigation, Conceptualization. **Dag Lindholm:** Writing – review & editing, Conceptualization. **Hallvard Fjær:** Writing – review & editing, Conceptualization. **Pernille Seljom:** Writing – review & editing, Methodology, Conceptualization. **Sean Erik Foss:** Writing – review & editing, Project administration, Funding acquisition, Conceptualization.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

Data will be made available on request.

Acknowledgments

This work was funded by the European Union's Horizon 2020 research and innovation program under grant agreement 818009 (Be-Smart).

The authors thank Dr. Heine Nygard Riise and Lisa Kvalbein for useful discussions.

References

[1] D.C. Jordan, T.J. Silverman, J.H. Wohlgemuth, S.R. Kurtz, K.T. VanSant, Photovoltaic failure and degradation modes, Prog. Photovoltaics Res. Appl. 25 (2017) 318–326, https://doi.org/10.1002/pip.2866.

[2] A. Skoczek, T. Sample, E.D. Dunlop, The results of performance measurements of field-aged crystalline silicon photovoltaic modules, Prog. Photovoltaics Res. Appl. 17 (2009) 227–240, https://doi.org/10.1002/pip.874.

[3] A. Virtuani, M. Caccivio, E. Annigoni, G. Friesen, D. Chianese, C. Ballif, T. Sample, 35 years of photovoltaics: analysis of the TISO-10-kW solar plant, lessons learnt in safety and performance—Part 1, Prog. Photovoltaics Res. Appl. 27 (2019) 328–339, https://doi.org/10.1002/pip.3104.

[4] D. Jordan, M. Kempe, I. Repins, J. Bleem, J. Menard, P. Davis, Life after 30 years- a PV system in Colorado, in: 2021 Photovoltaic Reliability Workshop, 2021.

[5] E. Annigoni, A. Virtuani, M. Caccivio, G. Friesen, D. Chianese, C. Ballif, 35 years of photovoltaics: analysis of the TISO-10-kW solar plant, lessons learnt in safety and performance—Part 2, Prog. Photovoltaics Res. Appl. 27 (2019) 760–778, https://doi.org/10.1002/pip.3146.

[6] D. Jordan, T. Barnes, N. Haegel, I. Repins, Build solar-energy systems to last — save billions, Nature 600 (2021) 215–217, https://doi.org/10.1038/d41586-021-03626-9.

[7] C.R. Osterwald, T.J. McMahon, History of accelerated and qualification testing of terrestrial photovoltaic modules: a literature review, Prog. Photovoltaics Res. Appl. 17 (2009) 11–33, https://doi.org/10.1002/pip.861.

[8] M. Owen-Bellini, S.L. Moffitt, A. Sinha, A.M. Maes, J.J. Meert, T. Karin, C. Takacs, D.R. Jenket, J.Y. Hartley, D.C. Miller, P. Hacke, L.T. Schelhas, Towards validation of combined-accelerated stress testing through failure analysis of polyamide-based photovoltaic backsheets, Sci. Rep. 11 (2021) (2019), https://doi.org/10.1038/s41598-021-81381-7.

[9] M. Owen-Bellini, P. Hacke, D.C. Miller, M.D. Kempe, S. Spataru, T. Tanahashi, S. Mitterhofer, M. Jankovec, M. Topič, Advancing reliability assessments of photovoltaic modules and materials using combined-accelerated stress testing, Prog. Photovoltaics Res. Appl. 29 (2021) 64–82, https://doi.org/10.1002/pip.3342.

[10] W. Gambogi, T. Felder, S. MacMaster, K. Roy-Choudhury, B.-L. Yu, K. Stika, H. Hu, N. Phillips, T.J. Trout, Sequential stress testing to predict photovoltaic module durability, in: 2018 IEEE 7th World Conference on Photovoltaic Energy Conversion (WCPEC) (A Joint Conference of 45th IEEE PVSC, 28th PVSEC & 34th EU PVSEC), IEEE, 2018, pp. 1593–1596, https://doi.org/10.1109/PVSC.2018.8547260.

[11] International Electrotechnical Commission, IEC TR 63279:2020- Derisking Photovoltaic Modules- Sequential and Combined Accelerated Stress Testing, 2020.

[12] G.C. Eder, Y. Voronko, S. Dimitriadis, K. Knöbl, G. Újvári, K.A. Berger, M. Halwachs, L. Neumaier, C. Hirschl, Climate specific accelerated ageing tests and evaluation of ageing induced electrical, physical, and chemical changes, Prog. Photovoltaics Res. Appl. 27 (2019) 934–949, https://doi.org/10.1002/pip.3090.

[13] M.D. Kempe, D. Holsapple, K. Whitfield, N. Shiradkar, Standards development for modules in high temperature micro-environments, Prog. Photovoltaics Res. Appl. 29 (2021) 445–460, https://doi.org/10.1002/pip.3389.

[14] D.C. Miller, J.G. Bokria, D.M. Burns, S. Fowler, X. Gu, P.L. Hacke, C.C. Honeker, M. D. Kempe, M. Köhl, N.H. Phillips, K.P. Scott, A. Singh, S. Suga, S. Watanabe, A. F. Zielnik, Degradation in photovoltaic encapsulant transmittance: results of the first PVQAT TG5 artificial weathering study, Prog. Photovoltaics Res. Appl. 27 (2019) 391–409, https://doi.org/10.1002/pip.3103.

[15] D.C. Miller, F. Alharbi, A. Andreas, J.G. Bokria, D.M. Burns, J. Bushong, X. Chen, D. Dietz, S. Fowler, X. Gu, A. Habte, C.C. Honeker, M.D. Kempe, H. Khonkar, M. Köhl, N.H. Phillips, J. Rivera, K.P. Scott, A. Singh, A.F. Zielnik, Degradation in photovoltaic encapsulation strength of attachment: results of the first PVQAT TG5 artificial weathering study, Prog. Photovoltaics Res. Appl. 28 (2020) 639–658, https://doi.org/10.1002/pip.3255.

[16] M.D. Kempe, T. Lockman, J. Morse, Development of testing methods to predict cracking in photovoltaic backsheets, in: 2019 IEEE 46th Photovoltaic Specialists Conference (PVSC), IEEE, 2019.

[17] K.-A. Weiss, L.S. Bruckman, R.H. French, G. Oreski, T. Tanahashi, Service Life Estimation for Photovoltaic Modules, 2021.

[18] S. Kurtz, K. Whitfield, N. Phillips, T. Sample, C. Monokroussos, E. Hsi, I. Repins, P. Hacke, D. Jordan, J. Wohlgemuth, P. Seidel, U. Jahn, M. Kempe, T. Tanahashi, Y. Chen, B. Jaeckel, M. Yamamichi, Qualification testing versus quantitative reliability testing of PV – gaining confidence in a rapidly changing technology, in: 33rd European Photovoltaic Solar Energy Conference and Exhibition, 2017, pp. 1302–1311.

[19] S. Kurtz, K. Whitfield, G. TamizhMani, M. Koehl, D. Miller, J. Joyce, J. Wohlgemuth, N. Bosco, M. Kempe, T. Zgonena, Evaluation of high-temperature exposure of photovoltaic modules, Prog. Photovoltaics Res. Appl. 19 (2011) 954–965, https://doi.org/10.1002/pip.1103.

[20] M.D. Kempe, J.H. Wohlgemuth, Evaluation of temperature and humidity on PV module component degradation, in: 2013 IEEE 39th Photovoltaic Specialists Conference (PVSC), IEEE, 2013, https://doi.org/10.1109/PVSC.2013.6744112, 0120–0125.

[21] M. Paggi, M. Corrado, M.A. Rodriguez, A multi-physics and multi-scale numerical approach to microcracking and power-loss in photovoltaic modules, Compos. Struct. 95 (2013) 630–638, https://doi.org/10.1016/j.compstruct.2012.08.014.

[22] T.J. Silverman, N. Bosco, M. Owen-Bellini, C. Libby, M.G. Deceglie, Millions of Small pressure cycles drive damage in cracked solar cells, IEEE J. Photovoltaics 12 (2022) 1090–1093, https://doi.org/10.1109/JPHOTOV.2022.3177139.

[23] J.H. Wohlgemuth, D.W. Cunningham, N.V. Placer, G.J. Kelly, A.M. Nguyen, The effect of cell thickness on module reliability, in: Proceedings of the IEEE PVSC, 2008.

[24] P. Lenarda, M. Paggi, A geometrical multi-scale numerical method for coupled hygro-thermo-mechanical problems in photovoltaic laminates, Comput. Mech. 57 (2016) 947–963, https://doi.org/10.1007/s00466-016-1271-5.

[25] P. Nivelle, J.A. Tsanakas, J. Poortmans, M. Daenen, Stress and strain within photovoltaic modules using the finite element method: a critical review, Renew. Sustain. Energy Rev. 145 (2021), 111022, https://doi.org/10.1016/j.rser.2021.111022.

[26] D. Lindholm, G. Otnes, H.G. Fjær, G. Cattaneo, H.Y. Li, S.E. Foss, Thermomechanical fatigue of solder joint and interconnect ribbon: a comparison between glass-glass and glass-foil modules, in: Proc. Of the 37th EUPVSEC, 2020.

[27] M. Springer, J. Hartley, N. Bosco, Multiscale modeling of shingled cell photovoltaic modules for reliability assessment of electrically conductive adhesive cell interconnects, IEEE J. Photovoltaics 11 (2021) 1040–1047, https://doi.org/10.1109/JPHOTOV.2021.3066302.

[28] J. Hartley, Multi-scale, Multi-Physics Modeling for PV Reliability, DuraMAT Webinar, 2019.

[29] M. Pander, U. Zeller, B. Jaeckel, M. Ebert, Digital prototyping- Application of numerical methods in module development, in: Proc. Of the 36th EUPVSEC, 2019.

[30] Q.Z. Zhang, B.F. Shu, M.B. Chen, Q.B. Liang, C. Fan, Z.Q. Feng, P.J. Verlinden, Numerical investigation on residual stress in photovoltaic laminates after lamination, J. Mech. Sci. Technol. 29 (2015) 655–662, https://doi.org/10.1007/s12206-015-0125-y.

[31] F. Kraemer, S. Wiese, E. Peter, J. Seib, Mechanical problems of novel back contact solar modules, Microelectron. Reliab. 53 (2013) 1095–1100, https://doi.org/10.1016/j.microrel.2013.02.019.

[32] M. Aßmus, K. Naumenko, H. Altenbach, A multiscale projection approach for the coupled global–local structural analysis of photovoltaic modules, Compos. Struct. 158 (2016) 340–358, https://doi.org/10.1016/j.compstruct.2016.09.036.

[33] N. Bosco, T.J. Silverman, S. Kurtz, The influence of PV module materials and design on solder joint thermal fatigue durability, IEEE J. Photovoltaics 6 (2016) 1407–1412, https://doi.org/10.1109/JPHOTOV.2016.2598255.

[34] F. Kraemer, S. Wiese, Assessment of long term reliability of photovoltaic glass–glass modules vs. glass-back sheet modules subjected to temperature cycles by FE-analysis, Microelectron. Reliab. 55 (2015) 716–721, https://doi.org/10.1016/j.microrel.2015.02.007.

[35] A.J. Beinert, P. Romer, M. Heinrich, M. Mittag, J. Aktaa, D.H. Neuhaus, The effect of cell and module dimensions on thermomechanical stress in PV modules, IEEE J. Photovoltaics 10 (2020) 70–77, https://doi.org/10.1109/JPHOTOV.2019.2949875.

[36] M. Pander, R. Meier, M. Sander, S. Dietrich, M. Ebert, Lifetime estimation for solar cell interconnectors, in: Proc. Of the 28th EUPVSEC, 2013.

[37] N. Bosco, T.J. Silverman, Solder bond fatigue is insensitive to module size, IEEE J. Photovoltaics 11 (2021) 1048–1050, https://doi.org/10.1109/JPHOTOV.2021.3074056.

[38] J. Zhu, M. Owen-Bellini, D. Montiel-Chicharro, T.R. Betts, R. Gottschalg, Effect of viscoelasticity of ethylene vinyl acetate encapsulants on photovoltaic module solder joint degradation due to thermomechanical fatigue, Jpn. J. Appl. Phys. 57 (2018), 08RG03, https://doi.org/10.7567/JJAP.57.08RG03.

[39] G. Li, M.W. Akram, Y. Jin, X. Chen, C. Zhu, A. Ahmad, R.H. Arshad, X. Zhao, Thermo-mechanical behavior assessment of smart wire connected and busbarPV modules during production, transportation, and subsequent field loading stages, Energy 168 (2019) 931–945, https://doi.org/10.1016/j.energy.2018.12.002.

[40] S. Kumar, R. Gupta, Investigation and analysis of thermo-mechanical degradation of fingers in a photovoltaic module under thermal cyclic stress conditions, Sol. Energy 174 (2018) 1044–1052, https://doi.org/10.1016/j.solener.2018.10.009.

[41] M.T. Zarmai, N.N. Ekere, C.F. Oduoza, E.H. Amalu, Optimization of thermo-mechanical reliability of solder joints in crystalline silicon solar cell assembly, Microelectron. Reliab. 59 (2016) 117–125, https://doi.org/10.1016/j.microrel.2015.12.031.

[42] International Electrotechnical Commission, IEC 61215 Series- Terrestrial Photovoltaic Modules- Design Qualification and Type Approval, International Electrotechnical Commission, 2021.

[43] N. Bosco, T.J. Silverman, S. Kurtz, Climate specific thermomechanical fatigue of flat plate photovoltaic module solder joints, Microelectron. Reliab. 62 (2016) 124–129, https://doi.org/10.1016/j.microrel.2016.03.024.

[44] M. Owen-Bellini, Thermomechanical Degradation Mechanisms of Silicon Photovoltaic Modules, Loughborough University, 2017.

[45] F.K.A. Nyarko, G. Takyi, E.H. Amalu, M.S. Adaramola, Generating temperature cycle profile from in-situ climatic condition for accurate prediction of thermo-mechanical degradation of c-Si photovoltaic module, Engineering Science and Technology, Int. J. 22 (2019) 502–514, https://doi.org/10.1016/j.jestch.2018.12.007.

[46] F.K. Afriyie Nyarko, G. Takyi, F.B. Effah, Study on creep damage in Sn60Pb40 and Sn3.8Ag0.7Cu (Lead-Free) solders in c-Si solar PV cell interconnections under in-situ thermal cycling in Ghana, Crystals 11 (2021) 441, https://doi.org/10.3390/cryst11040441.

[47] S.P. Aly, S. Ahzi, N. Barth, A. Abdallah, Numerical analysis of the reliability of photovoltaic modules based on the fatigue life of the copper interconnects, Sol. Energy 212 (2020) 152–168, https://doi.org/10.1016/j.solener.2020.10.021.

[48] M.U. Siddiqui, A.F.M. Arif, Electrical, thermal and structural performance of a cooled PV module: transient analysis using a multiphysics model, Appl. Energy 112 (2013) 300–312, https://doi.org/10.1016/j.apenergy.2013.06.030.

[49] K. Poncelet, H. Hoschle, E. Delarue, A. Virag, W. Drhaeseleer, Selecting representative days for capturing the implications of integrating intermittent renewables in generation expansion planning problems, IEEE Trans. Power Syst. 32 (2017) 1936–1948, https://doi.org/10.1109/TPWRS.2016.2596803.

[50] M. Hoffmann, L. Kotzur, D. Stolten, M. Robinius, A review on time series aggregation methods for energy system models, Energies 13 (2020) 641, https://doi.org/10.3390/en13030641.

[51] P. Seljom, L. Kvalbein, L. Hellemo, M. Kaut, M.M. Ortiz, Stochastic modelling of variable renewables in long-term energy models: dataset, scenario generation & quality of results, Energy 236 (2021), 121415, https://doi.org/10.1016/j.energy.2021.121415.

[52] M. Kaut, Scenario generation by selection from historical data, Comput. Manag. Sci. 18 (2021) 411–429, https://doi.org/10.1007/s10287-021-00399-4.

[53] A. Ghosal, A. Nandy, A.K. Das, S. Goswami, M. Panday, A short review on different clustering techniques and their applications, in: Emerging Technology in Modelling and Graphics, 2020, pp. 69–83, https://doi.org/10.1007/978-981-13-7403-6_9.

[54] IEA PVPS Task 13, Translation Tool for Geo Data to Koeppen-and-Geiger-Climate-Zone, (n.d.). https://iea-pvps.org/research-tasks/performance-operation-and-reliability-of-photovoltaic-systems/documents/..

[55] The Norwegian Meteorological Institute, Frost- Historical Weather Data, (n.d.). https://frost.met.no/..

[56] Deutscher Wetterdienst, CDC- Climate Data Center, (n.d.). https://cdc.dwd.de/portal/..

[57] U.S. Climate Reference Network, Quality Controlled Datasets, (n.d.). https://www.ncdc.noaa.gov/crn/qcdatasets.html..

[58] T. Huld, A. Amillo, Estimating PV module performance over large geographical regions: the role of irradiance, air temperature, wind speed and solar spectrum, Energies 8 (2015) 5159–5181, https://doi.org/10.3390/en8065159.

[59] W.F. Holmgren, C.W. Hansen, M.A. Mikofski, Pvlib python: a python package for modeling solar energy systems, J. Open Source Softw. 3 (2018) 884, https://doi.org/10.21105/joss.00884.

[60] D. Santos-Martin, S. Lemon, SoL – a PV generation model for grid integration analysis in distribution networks, Sol. Energy 120 (2015) 549–564, https://doi.org/10.1016/j.solener.2015.07.052.

[61] D.L. King, W.E. Boyson, J.A. Kratochvill, SANDIA REPORT-Photovoltaic Array Performance Model, 2004.

[62] Y. Liu, R. Sioshansi, A.J. Conejo, Hierarchical clustering to find representative operating periods for capacity-expansion modeling, IEEE Trans. Power Syst. 33 (2018) 3029–3039, https://doi.org/10.1109/TPWRS.2017.2746379.

[63] Scikit Learn, K-means clustering, (n.d.). https://scikit-learn.org/stable/modules/clustering.html#k-means (accessed September 5, 2023).

[64] Scikit Learn, Standard Scaler, (n.d.). https://scikit-learn.org/stable/modules/generated/sklearn.preprocessing.StandardScaler (accessed July 28, 2022).

[65] M.L. Rizzo, G.J. Székely, Energy Distance, vol. 8, Wiley Interdiscip Rev Comput Stat., 2016, pp. 27–38, https://doi.org/10.1002/wics.1375.

[66] dcor 0.5.6, Energy distance, (n.d.). , https://dcor.readthedocs.io/en/latest/functions/dcor.energy_distance.html (accessed September 5, 2023).

[67] International Electrotechnical Commission, IEC TS 63126:2020- Guidelines for Qualifying PV Modules, Components and Materials for Operation at High Temperatures, 2020.